Proceedings of the 4th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/83/2025.19941

A Mixture of Experts Approach for Refined Sarcasm Detection in Text

Zao Jiang[1,a,*]

[1]*Department of Computing, The Hong Kong Polytechnique University, Hong Kong, China*
a. zao311.jiang@connect.polyu.hk
**corresponding author*

Abstract: The topic of sarcasm detection has always been hot in natural language processing, which leverages the subtle contextual and linguistic cues of figurative language to identify and interpret it. Accurately detecting sarcasm is crucial for revealing the true sentiment expressed behind statements and thus to improve the effectiveness of sentiment analysis algorithms. This work introduces a new method of improving the accuracy of identifying sarcasm based on a Mixture of Experts scheme. The model is structured such that each expert network specializes in different aspects of sarcasm, captures distinct implicit features, and collectively detects sarcasm. Such structure enables a comprehensive understanding of context and sarcasm-related nuances. The model was evaluated on dialogue sarcasm datasets and achieved optimal performance. This improvement in detection performance is attributed to the model's ability to utilize expert specialization to better distinguish sarcastic content in various linguistic structures and contexts. It also shows promising results in terms of scalability to other datasets with less computational efforts and faster inference compared to larger and more complex language models, making this method a useful tool in other contexts social media monitoring and sentiment analysis. This study attempts to contribute to the further research in sarcasm detection and offer a foundation for future work on sentiment analysis.

Keywords: Mixture of Experts, Sarcasm Detection, Sentiment Analysis.

1. Introduction

Detecting sarcasm languages is a huge challenge in Natural Language Processing (NLP) since a sarcasm statement often has a meaning opposite to the used literal words; this makes it extremely hard to interpret the true intent behind a statement [1]. The challenge is significant in tasks such as social media monitoring, dialogue systems, and sentiment analysis, where understanding user intent is crucial. For instance, a sarcastic statement might seem positive based on its surface sentiment, but the underlying intent could be negative. Additionally, if sentiment analysis systems misinterpret sarcasm, they may provide awkward or inappropriate responses that negatively affects effectiveness and user satisfaction of these systems.

The task of identifying sarcasm is inherently challenging [2,5]. Sarcasm is heavily context-dependent, which requires a nuanced understanding of textual surroundings and conversational and social contexts. Also, sarcasm typically expresses the opposite of what is literally said. The problems created by this reversal of meaning also pose difficulties for traditional models reliant upon surface

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/83/2025.19941

level sentiment indicators. In addition, sarcasm often lacks explicit linguistic markers, so the identification process is without deeper contextual understanding. Unlike the possibly more identifiable patterns which humor or exaggeration have, sarcasm's implicit features are more complex for machine learning models to capture. Lastly, sarcasm appears relatively infrequent in many datasets. This data imbalance problem may exacerbate the difficulty of training robust models capable of detecting it.

Traditional approaches to sarcasm detection have primarily relied on early machine learning models like Logistic Regression and Support Vector Machine (SVM). Previous work used handcrafted features consisting lexical patterns, syntactic structures, and metadata like user profiles or context-specific information to identify sarcastic expressions [6]. These models were effective to some extent but struggled to handle the inherent complexity and scope of sarcastic language, especially when contextual clues or tone were in play. Furthermore, these methods had to be more generalizable across a wider range of linguistic context. The transformer models, especially Bidirectional Encoder Representations from Transformers (BERT)-based models, have shown promise by leveraging bidirectional context and long-range dependencies [3]. Despite the improvements in understanding the context, these models face significant limitations regarding sarcasm detection. The limitations include an oversimplification of sarcasm as a binary classification task, sparse representation of sarcastic examples in general-purpose datasets, and difficulty fully incorporating broader contextual knowledge, such as tone or speaker intent, which are critical for recognizing sarcastic expressions and statements.

In this study, we introduce a novel method of sarcasm detection based on the Mixture of Experts (MoE) model [4,13]. To addresses the limitations of existing methods by combining multiple BERT-based expert networks, each specializing in a different aspect of sarcasm detection and employing a dynamic gating mechanism to route input samples to the most appropriate expert. This specialization allows the model to handle various challenges in distinguishing sarcasm more effectively. The contributions of this paper include: (1) Provided analysis on existing sarcasm detection methods, ranging from early machine learning models to transformer-based methods. (2) Proposed a sarcasm detection method using the MoE structure, which provides a solution to the challenges of sarcasm detection by integrating specialized experts. (3) Demonstrated the effectiveness of the MoE based method through experiments, which offers improved F1-score in detecting sarcastic expressions.

2. Related Work

Sentiment analysis has become an area of abound in the broad field of NLP and identifying sarcasm is a subtask of it. Sarcasm often involves a contradiction between literal and intended meanings, making traditional NLP models struggles to interpret. Effective sarcasm detection is crucial for accurately analyzing sentiment and user intent, especially in social media which is dominated by sarcasm. This section explores and discusses various methods in sarcasm detection, providing insights into the existing methods.

Traditional machine learning models, namely SVM and Logistic Regression, are primarily used for the early work on sarcasm detection [5]. These approaches were practically relying on the handcrafted features, such as lexical, syntactic, semantic pattern, and user and context specific information. However, these methods struggled to capture the complexities and nuances of sarcastic language, especially in cases where contextual or tonal cues were necessary. Bamman et al. introduced a binary logistic regression model for detecting sarcasm on Twitter by incorporating features from the tweet, the author, the audience, and the environmental context [6]. The method relies upon manually extracted features and achieves good performance. This also makes it lack the ability to generalize across broader linguistic contexts. To address data imbalance issues, Liu et al. proposed an ensemble learning approach, MSELA, which integrates multiple strategies to better

DOI: 10.54254/2753-8818/83/2025.19941

handle the class imbalance [7]. Their method incorporated feature selection and boosting techniques, though it faced challenges capturing deeper contextual dependencies inherent in sarcasm.

As the limitations of feature-based methods became apparent, the field started shifting towards deep learning approaches. Models including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) beat traditional methods by learning more intricate patterns and relations within text data. For instance, CNNs perform great at identifying local patterns, whereas RNNs are more suitable for handling sequential data. These models, on the other hand, had notable limitations. CNNs and RNNs often struggled with long-range dependencies and the complex contextual understanding required for identifying sarcasm. Ghosh et al. addressed some of these shortcomings by combining CNNs and Long Short-Term Memory (LSTM) networks, achieving strong performance in text-based sarcasm detection [8]. However, the complexities of sarcastic expression prevented such architecture from achieving deeper contextual understanding.

The advancement of transformer-based models like BERT and GPT significantly improved sarcasm detection by utilizing attention mechanisms. These models outperform at capturing contextual information and long-range dependencies, which is important to understanding sarcasm. The attention mechanisms enables the model to focus on relevant input segments when making predictions, thus making more accurate interpretations of sarcasm expressions. Building on this, Li et al. introduced a context-aware, attention-based multimodal fusion method for sarcasm detection to model emotional mismatches between modalities [9]. Their work introduced novel mechanisms for detecting inter-modal emotional inconsistency and contextual scenario mismatches, highlighting how cross-modal attention can capture more nuanced features. By leveraging multimodal data and complex contextual information, these advances shift the sarcasm detection to a completely new level. Hiremath et al.'s method also utilizes multimodal features. Their model captures cognitive features of voice cues, eye movements, and linguistic features and presented a high potential for sarcasm detection [10]. Tiwari et al. developed the method of Quantum Fuzzy Neural Network (QFNN) that integrates neural networks with fuzzy logic to perform multimodal sentiment and sarcasm detection [11]. Their approach combines linguistic and non-linguistic cues to improve the performance of distinguishing sarcasm, particularly in multimodal data. Zhang et al. combined BERT with a stance-centered graph attention network (SCGAT) to learn from the sentence representation and specific target's stance information, proving the capacity of the proposed framework in detecting sarcasm [12].

The existing models have shown great promise but are mostly concentrating on longer and contextual rich texts or multimodal data. As opposed to extracting sarcasm from contextual dense texts, our method specializes in short text context-less sarcasm detection, a challenging but highly relevant domain in modern social media communication. By employing a MoE model, we aim to handle the variability of sarcastic expressions efficiently. In short text scenarios, the approach leverages expert networks that specialize in various aspects of sarcasm, which allows better feature representation. Thus, the model’s performance is comparable to large language models (LLMs) but with better computational efficiency, making the model more scalable and practical for real-world applications. This study adds to the previous work, focusing on the challenges of short-text sarcasm detection and contributes a novel and efficient solution to the field.

3. Method

In this study, a MoE model is employed for sarcasm detection in posts on the internet by explicitly integrating diverse expert knowledge into a collective decision-making process. This approach provides an adaptive and scalable solution to the challenging task of sarcasm detection, where the meaning of sarcasm can vary significantly across texts.

DOI: 10.54254/2753-8818/83/2025.19941

3.1. MoE Architecture

Figure 1 shows the model's structure, based on the MoE framework consisting of the expert networks and a gating network. The MoE layers are sparse layers used to replace the feed-forward network (FFN) layers [14]. We have employed four pre-trained BERT or RoBERTa models as the experts to cover a comprehensive range of sarcasm detection capabilities in this approach. The goal of training the MoE model is to have each expert specialize in capturing distinct features of the sarcastic language and its understanding of sarcasm. The process enables each expert to use the features that are most discriminative to distinguish sarcasm.

The reason for using BERT-based models is its ability to model complex linguistic features, particularly in natural language understanding. BERT-based methods use a bidirectional attention mechanism to consider the preceding and succeeding words when encoding sentences [15]. This bidirectional processing enables BERT to capture the subtle changes in tone, intent, and context that are often critical to identify sarcasm. It has shown in various works that BERT-based models' better performance in sarcasm detection [3] [14]. To identify sarcasm, both literal meaning and implied meaning is needed. BERT-based methods' ability to capture intricate contextual cues makes it an ideal candidate for this task. Moreover, its fine-tuning capability allows it to adapt to specific datasets, making each expert network in the MoE to specialize in specific features of sarcastic expressions without requiring architectural modifications for tasks.

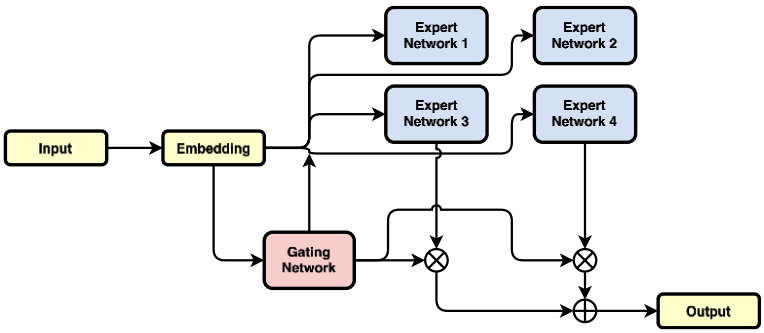

Figure 1: Proposed MoE model architecture for sarcasm detection

3.2. Gating Mechanism

Another important component of MoE is a gating network. It determines which and how input tokens are sent to the expert network. The gating network is trained to be a dynamic selection system that decides the relevance of each expert for a given input so that the outputs from most relevant experts are combined to provide the final output.

The general MoE structure contains multiple expert functions, denoted as $f_1, f_2, \ldots, f_n$ where each expert function f_i takes the same input x and generates output $f_i(x)$. The gating function w also takes x as input and produces a vector of outputs of the weights assigned to the experts: $w(x) = (w_1(x), w_2(x), \ldots, w_n(x))$. We represent the vector of parameters by $\theta = (\theta_0, \theta_1, \ldots, \theta_n)$, where θ_0 being the gating function's parameters. For the given input x, the model combines the outputs of all experts with corresponding and produce an aggregated output. This allows the model to integrate diverse expertise dynamically based on the input.

The loss function used are adapted from router Z-loss in Zoph et al.'s work St-moe, which could improve training stability of MoE without quality degradation [17]. The router Z-loss is works by penalizing large logits entering the gating network, which can be formulated as:

$$L_z(x) = \frac{1}{B}\sum_{i=1}^{B}\left(\log\sum_{j=1}^{N} e^{x_j^{(i)}}\right)^2 \quad (1)$$

DOI: 10.54254/2753-8818/83/2025.19941

where N represents the number of experts and B is the number of tokens. It is capable to stabilize the model without any performance degradation. The router Z-loss reduces the roundoff error by controlling the gating function and ensuring small absolute magnitude of parameters. The total loss used in our MoE follows Zoph et al.'s paper [17]: a combination of the cross-entropy loss, the auxiliary load balance loss, and the router Z-loss, written as:

$$L = L_{CE} + c_B L_B + c_z L_Z \tag{2}$$

where L_{CE} stands for the cross-entropy loss and L_B stands for the load balance loss, with corresponding weights. With the combined loss, we can optimize the stabilized model without affecting model quality.

4. Experiment

4.1. Experiment Setup

For the evaluation of the proposed MoE model, the Sarcasm Corpus V2 is used, which is a subset of the Internet Argument Corpus (IAC) specifically annotated for sarcasm [18]. The dataset extends Sarcasm Corpus V1 and includes three categories of sarcasm: generic sarcasm (GEN), hyperbole (HYP), and rhetorical questions (RQ). Each class of the data has 6,520, 582, and 851 posts, respectively, with balanced sarcastic and non-sarcastic classes. HYP posts include trigger words indicating exaggeration, while RQ posts feature question-response pairs in which the speaker carries on with their part.

To be noted that the approach does not train the MoE model's gating algorithm to recognize a specific class of sarcasm for each expert network explicitly. Instead, we allow the gating mechanism to autonomously determine which input features are most relevant to each expert, promoting an implicit specialization that may better capture the nuanced and varied nature of sarcasm across different contexts. This strategy avoids overfitting predefined sarcasm types, enhancing the model's generalizability.

4.2. Model Training and Evaluation

The training process in this work involved fine-tuning pre-trained BERT-based expert networks within the MoE framework. The number of experts is chosen to be 4 as it already shows promising results on the dataset. The model is trained for 5 epochs using a single A-100 GPU. The gating algorithm is trained simultaneously using the router Z-loss to ensure classification accuracy and model stability. To provide a robust evaluation, the training is done by 5-fold cross-validation with 20% of data held out as testing data, following the scheme of previous work done by Jang et al. [16].

As the MoE structures tend to overfit the training set, early stopping was incorporated to prevent overfitting [19]. While keeping track of the training and validation accuracy, the drop happens after the third epoch. The classification results are measured using the average precision, recall, and F1 score. The model is compared with several baselines, including a vanilla SVM, a Logistic Regression model, three BERT-based models by Jang et al., and a multi-head self-attention with gated recurrent units (GRU) approach by Akula et al. [16,20]. The BERT-based models include the original BERT, RoBERTa, and DeBERTa. Akula et al. did not state if the evaluation is done under the 5-fold cross-validation scheme, and Jang et al. did not provide the precision and recall for the BERT-based models.

4.3. Results and Discussion

As shown in Table 1 and Table 2, our experiments yielded promising results, demonstrating the effectiveness of the MoE model in handling the complexity of sarcasm detection across multiple

DOI: 10.54254/2753-8818/83/2025.19941

categories. The average precision, recall, and F1-scores of proposed BERT-based models under MoE framework remain the highest among other solely BERT-based and multihead-self attention with GRU methods. The model can effectively recognize sarcastic language and statements, suggesting the MoE framework can capture the implicit sarcastic expressions.

The MoE model outperformed the baselines, which underscores the value of using a MoE scheme, especially in tasks requiring the detection of subtle and context-dependent linguistic phenomena like sarcasm. This is done by letting each expert learn and specialize in a subset of tokens or implicit features, and collectively create an efficient model. Another finding is that the MoE model did not require too much training and computational resources compared with LLMs or other complex BERT-based models. The four expert MoE models could provide comparable results with the state-of-the-art models. Notedly, the dataset used is a short-text context-less sarcasm dataset, while some LLMs might achieve better results on context-dependent sarcasm detection and outperform the non-LLM models. It is quite promising that the MoE structure could provide good results and does not require tremendous training as the model size is relatively small. This ensures the scalability of such models and applications on various tasks in sentiment analysis.

In summary, the experimental results demonstrate that our MoE model, leveraging the advanced capabilities of BERT-based models within a flexible expert-gating framework, offers a powerful approach to sarcasm detection. The model's capability to dynamically adapt and specialize from the input data allows it to capture the diverse and nuanced nature of sarcastic expressions more effectively than traditional methods.

Table 1: Sarcasm Detection Performance on Sarcasm Corpus V2.

Method	Precision	Recall	F1-Score
SVM	0.74	0.74	0.74
Logistic Regression	0.73	0.73	0.72
BERT [16]	/	/	0.77
RoBERTa [16]	/	/	0.80
DeBERTa [16]	/	/	0.78
Multihead-Self Attention + GRU [20]	0.77	0.77	0.77
MoE with BERT	0.80	0.80	0.80
MoE with RoBERTa	**0.83**	**0.82**	**0.82**

Table 2: Confusion Matrix of the MoE models

True Labels	Predicted Labels (MoE with BERT)		Predicted Labels (MoE with RoBERTa)	
	Non-Sarcastic	Sarcastic	Non-Sarcastic	Sarcastic
Non-Sarcastic	716	223	692	247
Sarcastic	160	779	90	849

5. Conclusion

In this paper, a sarcasm detection method is designed based on the MoE model. Extensive experiments demonstrated the effectiveness of the MoE method for sarcasm detection. In particular, the MoE model achieves better classification accuracy in handling various types of sarcasm compared to previous solely BERT-based methods. It suggests that the MoE can dynamically allocate expert networks based on the input tokens via a router network, even when the sarcasm cues are subtle. This paper highlights the potential of using architectures like MoE for sarcasm detection in NLP and can be extended to other sentiment analysis tasks. The strong performance and efficiency of the MoE model in detecting generic sarcasm underscore the advantages of a modular approach, where different

DOI: 10.54254/2753-8818/83/2025.19941

expert networks can specialize in various aspects of sarcasm detection. This provides a significant advancement over the generalized models, which may overlook the subtle marks and context-dependent nature of sarcastic languages. However, sarcasm detection and sentiment analysis still remain a challenging and evolving task in NLP.

The study demonstrates the potential of MoE structures to capture the intricate and varied forms of sarcastic language. Future work could focus on enhancing the MoE structure as the proposed method is relatively simple or could utilize more contextual information to achieve better performance.

References

[1] *Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic sarcasm detection: A survey. ACM Computing Surveys (CSUR), 50(5), 1-22.*

[2] *Chaudhari, P., & Chandankhede, C. (2017). Literature survey of sarcasm detection. In 2017 International conference on wireless communications, signal processing and networking (WiSPNET) (pp. 2041-2046). IEEE.*

[3] *Baruah, A., Das, K., Barbhuiya, F., & Dey, K. (2020). Context-aware sarcasm detection using BERT. In Proceedings of the Second Workshop on Figurative Language Processing, 83-87.*

[4] *Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural computation, 3(1), 79-87.*

[5] *Moores, B., & Mago, V. (2022). A survey on automated sarcasm detection on Twitter. arXiv preprint arXiv:2202.02516.*

[6] *Bamman, D., & Smith, N. (2015). Contextualized sarcasm detection on twitter. In proceedings of the international AAAI conference on web and social media (Vol. 9, No. 1, pp. 574-577).*

[7] *Liu, P., Chen, W., Ou, G., Wang, T., Yang, D., & Lei, K. (2014). Sarcasm detection in social media based on imbalanced classification. In Web-Age Information Management: 15th International Conference, WAIM 2014, Macau, China, June 16-18, 2014. Proceedings 15 (pp. 459-471). Springer International Publishing.*

[8] *Ghosh, A., & Veale, T. (2016, June). Fracking sarcasm using neural network. In Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 161-169).*

[9] *Li, Y., Li, Y., Zhang, S., Liu, G., Chen, Y., Shang, R., & Jiao, L. (2024). An attention-based, context-aware multimodal fusion method for sarcasm detection using inter-modality inconsistency. Knowledge-Based Systems, 287, 111457.*

[10] *Hiremath, B. N., & Patil, M. M. (2021). Sarcasm detection using cognitive features of visual data by learning model. Expert Systems with Applications, 184, 115476.*

[11] *Tiwari, P., Zhang, L., Qu, Z., & Muhammad, G. (2024). Quantum fuzzy neural network for multimodal sentiment and sarcasm detection. Information Fusion, 103, 102085.*

[12] *Zhang, Y., Ma, D., Tiwari, P., Zhang, C., Masud, M., Shorfuzzaman, M., & Song, D. (2023). Stance-level sarcasm detection with bert and stance-centered graph attention networks. ACM Transactions on Internet Technology, 23(2), 1-21.*

[13] *Eigen, D., Ranzato, M. A., & Sutskever, I. (2013). Learning factored representations in a deep mixture of experts. arXiv preprint arXiv:1312.4314.*

[14] *Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538.*

[15] *Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.*

[16] *Jang, H., & Frassinelli, D. (2024). Generalizable Sarcasm Detection Is Just Around The Corner, Of Course!. arXiv preprint arXiv:2404.06357.*

[17] *Zoph, B., Bello, I., Kumar, S., Du, N., Huang, Y., Dean, J., ... & Fedus, W. (2022). St-moe: Designing stable and transferable sparse expert models. arXiv preprint arXiv:2202.08906.*

[18] *Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., & Walker, M. (2017). Creating and characterizing a diverse corpus of sarcasm in dialogue. In The 17th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL).*

[19] *Xie, Y., Huang, S., Chen, T., & Wei, F. (2023, June). Moec: Mixture of expert clusters. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 11, pp. 13807-13815).*

[20] *Akula, R., & Garibay, I. (2021, April). Explainable detection of sarcasm in social media. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 34-39).*